

60409684.091002

SYSTEM AND METHOD FOR CONSOLIDATED DATA STORAGE
AND DATA PROTECTION

BACKGROUND OF THE INVENTION—FIELD OF INVENTION

This invention is directed to data storage and protection techniques, and in particular, to an improved distributed system for storing and protecting computer-based data.

BACKGROUND OF THE INVENTION

Businesses are seeing an ever-increasing demand for storing data in digital form. The methods for storing, protecting and managing data have not changed significantly in 40 years. Primary data is stored on magnetic disk drives and data is protected using backup software, systems and magnetic tapes. Storage administrators are required to assemble, integrate, test and debug a dozen individual data storage and data protection "components" to build a complete data storage and data protection system. This creates interoperability problems among these individual components, leaving the customer to resolve these issues out with their respective storage vendors. A system that would more tightly integrate data storage and data protection components would eliminate these interoperability issues.

Many new applications are demanding storage systems that can support tens to hundreds of terabytes of capacity. Data storage vendors offer storage solutions whose capacity is an order of magnitude less than the demand. This forces storage administrators to deploy dozens of storage components to achieve the storage capacity required by the application. Unfortunately, even when dozens of these similar components are deployed, they still do not behave and operate as one large storage capacity pool. Instead they act as many individual pools of capacity. Each pool is filled with data by applications that create data at different rates of growth. This causes some of these storage components to be filled with data while others have very little consumed capacity, resulting in an overall utilization of ~50%. In order to increase utilization today, storage administrators are responsible for migrating data from full storage components to empty or less-full storage components. This activity is time consuming and an error-prone process. A system that would seamlessly scale in capacity

60409684.091002

well beyond the needs of today's demand would significantly reduce the administrative issues associated with today's products.

Today, most data storage systems are deployed to serve the needs of applications that operate from the same site. The data storage industry has created very few products that allow storage to span two or more sites. The products that do exist are vendor specific, costly, proprietary solutions. Because of these limitations, there has been very low acceptance of these solutions. Also, the administrative burden of maintaining multiple storage and server systems at multiple sites is significant because tools have not been developed to allow administrators to manage multi-site deployments.

There are block, file and database replication products available today that perform multi-site replication of data. These replication products do not eliminate the need for regular tape backup processes. With file replication system in place, if a user or application accidentally deletes one copy of a file, all associated replicas are deleted as well.

A product that more effectively ties together enterprise site "islands" of storage capacity and performs more intelligent replication would thus be desirable. It would also serve to raise the overall utilization of these storage resources, thereby saving money by deferring the purchase of new storage systems.

Data stored on magnetic disk drives must be protected from being lost. Data loss can occur when a user or application accidentally deletes data, or when a disk drive or disk subsystem component fails. There are a number of data protection schemes that are employed today. Each of these will be discussed along with their shortcomings:

- (1) With RAID disk subsystems, data is redundantly stored on two or more disk drives. RAID protects data if from the loss of a single disk drive. In the event of a single drive failure, one or more of the surviving disk drives can continue to deliver data. These systems are relatively expensive since all of the support components must be duplicated; power supplies, controllers, internal buses and the drives themselves. These systems also cannot deliver data if more than a single drive fails in a set of redundant drives. On most RAID subsystems, when a disk drive fails, it must be replaced with the exact same drive make/model and storage capacity, which can be challenging

60409684.091002

when drive technology advances as quickly as it does. The first failed disk drive in a RAID set must be replaced within minutes or hours of the failure to minimize the exposure of data loss that would result if a second drive were to fail in the RAID set. When two or more drives have failed in a typical RAID set, the data on the surviving drives is effectively useless. A recovery from the latest set of backup tapes is the only mechanism for recovering from a multiple drive failure within a RAID set. The administrative overhead associated with managing a RAID subsystem is significant, and often involves the management of many other components such as storage area network (SAN) switches, host bus adapters, volume management and file system software, alternate path software, and at least one associated management tool for each storage component.

- (2) Host-based volume managers can also be used to protect data by replicating blocks (a block represents one or more 512 byte disk drive sectors) across multiple volumes that are presented by RAID disk subsystems. Volume management is typically administered by a systems administrator and RAID systems are typically managed by a storage administrator. One common problem with the popular combination of host-based volume management and RAID disk subsystems is the over-protection of data. If redundancy is applied using both the volume management system and the RAID storage system, customers would have their data doubly-replicated. This is a common occurrence because the system administrator and the storage administrator did not communicate the configuration of the volumes to each other. Today, there are no effective storage management tools to detect and remedy these disk volume configuration errors. RAID and volume management are crude tools for managing disk storage capacity.
- (3) By far, the most common form of magnetic disk data protection is magnetic tape backup. In this model, data from the magnetic disk subsystem is replicated on a regularly scheduled (typically daily) basis to magnetic tape by using backup software and one or more tape drives and tape media. Magnetic tape technology and today's backup software offer a less-than-ideal data protection solution:

60409684.091002

- a. Magnetic tape technology has not been able to maintain the same cost-per-gigabyte reductions that magnetic disk drives have experienced over the past 17 years. In 1985, the cost/gigabyte of magnetic tape media was 1/100th the cost of magnetic disk drives, making it a cost-effective choice as a backup medium. In 2002, magnetic disk drives and magnetic tape media are reaching cost parity at approximately -\$1/gigabyte. If one considers the additional administrative costs associated with managing magnetic tape, such as mechanical tape library units and complicated tape drives, it's easy to demonstrate that magnetic tape is already a much more costly data protection solution than a system built on just magnetic disk technology. If this 17 year cost / capacity trend continues, and there's no reason to assume it will not, it will cost more to store data on magnetic tape than the magnetic disk drives that the tapes are trying to protect. The existing data storage product vendors have not responded effectively to this relative cost trend, and continue to perpetuate the same tape-based backup model for data protection.
- b. Magnetic tape reliability has always been poor. When tape media contacts a tape drive head, both the tape media and the tape drive read/write heads are worn down due to the direct contact they have with each other. Magnetic disk drives have no head to media contact during normal operation.
- c. The operating and archive environmental requirements for magnetic tape are much more restrictive than magnetic disk.

| | Magnetic Tape | Magnetic Disk |
|------------------------|-----------------|---------------------|
| Operating Temp (C) | 10-40 degrees C | 5-55 degrees C |
| Operating Humidity (%) | 20-80% | 5-95% |
| Archive Temp (C) | 18-28 degrees C | -40 to 65 degrees C |
| Archive Humidity (%) | 40-60% | 5-95% |

- d. Tape media recording quality diminishes over time, forcing customers to refresh their tapes periodically. This involves copying data from the older tape onto a new tape. Since this is a time-intensive process, most customers just accept that there will be regular occurrences of unreadable

60409634.091002

tapes. A data protection system that automatically and periodically detects and corrects all instances of unreadable data on all storage media would be highly marketable.

- e. There are many incompatible magnetic tape and tape drive technologies. Even within a product line from a single vendor, there are older versions of tape that are not readable by the latest version of that vendor's tape drive. Once a technology is selected by a customer and is used for many years, it's extremely difficult to change to a different tape technology. Large repositories of tapes either have to be migrated to the new tape technology or the administrator must maintain two incompatible tape drive systems. Magnetic disk drive technology eliminates these problems. Even though there are a wide variety of magnetic disk drives to choose from in terms of form-factor, cost, performance, and capacity, the technology can be upgraded or changed with no significant administrative issues. A data protection system that replaces all magnetic tape technology with magnetic disk would provide greatly simplified administration.
- f. Removable data storage media like magnetic tape and optical disks (CD-R, DVD, WORM and magneto-optical disks) create administrative overhead in managing the physical location of media. Media is moved from online to offline and possibly to offsite locations in order to protect from site disaster. A protection system that maintains a local and a remote copy of data to protect from site disaster and also maintains a history of all versions of data in multiple locations would be highly marketable.
- g. Offsite storage of backup media at a third party vaulting company is expensive. Products that can leverage the disk storage capacity that is currently unused within other sites of their own enterprise in order to protect their data would be highly marketable.
- h. When magnetic tape must be used to recover data after the failure of an entire computer system or loss of an entire site, the entire recovery process of can take days or even

60409684.091002

weeks. It involves locating the right collection of tapes, placing these into tape library units and serially pulling data from tape and writing it to new disks. A system that continually maintains reliable disk-based replicas of all data for the purpose of providing instant site disaster recovery would be highly marketable.

- i. A LAN administrator (reported in eWeek magazine 8/26/2002) reported his experiences with magnetic tape reliability: "In any given 12 month period, about half of my tapes broke and had to be replaced. And whenever I needed to restore from a backup tape, the restore failed about half the time, even with every conceivable error-checking option turned on." Statistics reported by administrators on the quality of tape media run the gamut, from being very reliable to being abysmally unreliable. A system that eliminates this variability in the quality of the storage media would be highly marketable.
- j. Backup software is a major contributor to wasted media cost due to over-replication of data. Each week, most companies perform full backups and maintain as much as a years worth of these full backups. Typically each full backup will contain 80% of the same content as the last full backup. So after a year, the customer has over 50 sets of replicated data. A protection system that conservatively consumes storage resources, yet provides full recovery from the loss of one or more files, disk drives, storage systems or entire sites would be highly marketable.
- k. It is difficult for an administrator to know which tapes can be eliminated from a large tape archive. The critical content that must be preserved is commingled on the same tape medium with content that could be deleted. For this reason, tape archives continue to grow ad-infinity. Also, there is no software that reports on how long data is being maintained in an archive, making regulatory compliance checking impossible. A product that can completely manage the full lifecycle of data from creation to deletion would be desirable.

60409684.091002

1. It is impossible to ascertain the quality of data on specific magnetic tapes within an archive without placing each medium into a tape drive and reading that medium from beginning to end. The act of performing this tape integrity checking itself creates additional tape media and drive head wear. It takes hours to complete the scan of one tape alone. Even if this process were not so time consuming, when data is found to be damaged, there is no way to repair the tape to restore the data to its proper readable form. A storage and protection system that continually performs integrity checking on all data and can successfully correct any data that is damaged automatically would be a highly marketable product.
 - m. Backup software scheduling is too simplistic a protection model for capturing changes to data in a data storage system. At some time interval, typically once a day, the state of the disk system is taken as a "snapshot" and copied to tape. If multiple changes were made to a data file between two backup runs, all of the intermediate versions would not be protected. Snapshot products have similar limitations and are no replacement for traditional backups. A product that could capture all changes to files and allow customers to view the state of their environment at any point in time would be desirable.
- (4) A few new storage product vendors are leveraging disk subsystems as part of a backup protection scheme.
- a. Some vendors have developed "virtual tape" products that treat a magnetic disk subsystem as a cache in front of a magnetic tape library unit. This increases the "component" count of the storage system which in turn greatly increases the administration overhead with managing multiple levels of backup data. It also does not eliminate all of the aforementioned problems with the cost, reliability and management aspects of magnetic tapes.
 - b. Some vendors have designed disk subsystems that look to backup software like tape media. This solution wastes magnetic disk capacity because every full backup contains about 80% the same content as the full backup before it.

60409684.091002

c. Some vendors support a large magnetic disk repository for holding point-in-time snapshots of data. These snapshots do not eliminate traditional backups since they cannot be maintained for extended periods of time.

Finally, the management of data storage and data protection systems today suffers from the following problems:

- (1) Each storage component within a storage system is complex. If it takes a dozen components to completely build a storage system, it takes storage administrators, system administrators, backup administrators, network administrators and disaster recovery administrators to fully manage the entire storage system. It is often the case that these administrators do not communicate each of their configuration changes to each other, leading to problems like loss of data or over-replication.
- (2) The lack of common terminology across products with similar functionality (e.g. backup), even within a single vendor's family of products, creates vendor/product administrative specialization. A backup administrator may only be skilled in one product from one vendor since all other backup products use different terminology, and have very different configuration and administration processes.
- (3) A number of companies are attempting to develop "super storage management" applications that support all data storage and protection hardware and software components and revisions. It is doubtful that anyone can succeed at building and maintaining such a product.
- (4) Hardware and software products have been serviced in a 24x7 reactive mode model for the past 40 years. A vendor that can provide self-healing, auto-recovery of failed hardware will have a marketable edge against these traditional systems.

SUMMARY OF THE INVENTION

The following is a summary of the objects and advantages of the present invention which is directed to new and improved data storage, data protection and data management techniques.

- (1) In the data storage area, the present invention provides:

60409684.091002

- a. A shareable storage system that allows multiple sites to cooperate in sharing their storage resources for both data storage and data protection.
- b. A scalable storage system that can be expanded to petabytes in increments of one or more terabytes. It leverages grid-computing technology to deliver this scale across hundreds of enterprise sites.
- c. A reliable storage system that provides access to all data all the time. When nodes of the system are not available, surviving nodes continue to deliver data to requesting applications.
- d. A high performance storage system that provides the same access time to all data, regardless of when it was last stored. This is a marked improvement over current archiving technology where it may take days to acquire a file from a tape stored in an offsite location.
- e. A flexible system that allows independent expansion of storage capacity and storage bandwidth.
- f. A virtualized storage system that delivers location-independent data storage. Data is automatically written to any storage resource that has available capacity within the specified site repositories. Location-independence eliminates the complex administration of host-based volume management, SAN management, and RAID disk subsystem management.
- g. A highly available storage system that can continue to deliver data even when components fail. Unlike most of today's data storage products, no special purpose mechanical hardware such as hot-swap power supplies and disk drives are required.
- h. A capacity-conservative storage system that maintains versions of data for every change of a file. Optionally, older versions of files can be delta-compressed by identifying the differences between two versions and compressing these deltas.
- i. A cost-effective storage system that is built from high-volume server hardware offered by major computer systems vendors (e.g. Dell, HP, IBM).

60409684.091002

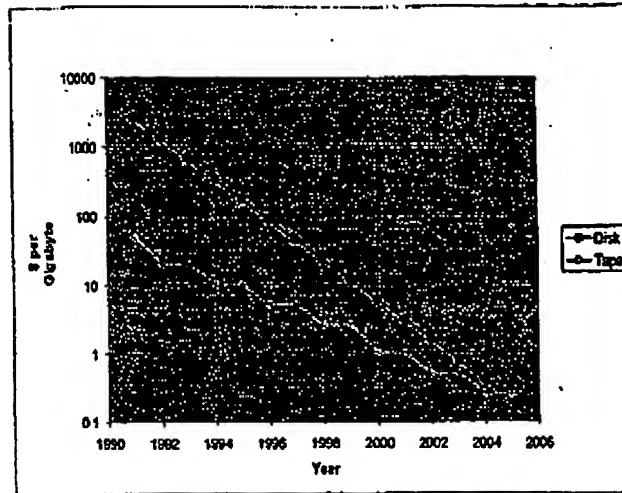
- j. A future-proof data storage system that directly leverages the advances of magnetic disk technology to provide increased storage density. When storage units must be replaced after years of operation, they can be replaced with units that provide much more capacity within the same physical space.
- (2) In the area of data protection, the present invention provides:
- a. A consolidated data protection system that replaces backup, archiving and replication software with a single consolidated data protection scheme.
 - b. A quick-recovery data protection system that greatly reduces the amount of time that it takes to recover from a system failure or site disaster, since all data is already resident on magnetic disks at surviving sites.
 - c. A responsive data protection system that protects all changes to new or updated files. Conversely, no additional storage capacity is consumed for data that has not changed. This is an improvement over traditional snapshot and backup techniques that only provide a limited point-in-time view of data files and blindly replicate content even when it has not changed.
 - d. A simpler data protection system that replaces magnetic tape as the primary means of protecting magnetic disk data with a system that leverages other disks and networks to provide a greater level of protection.
 - e. A full-lifecycle data protection system that manages the entire lifecycle of a file from creation to deletion.
 - f. A consolidated data protection system that allows an offsite copy of data to be used as a replacement for traditional offsite backup/archive tape storage as well as for site disaster recovery.
 - g. A flexible and complete data protection system that allows an administrator to define a policy on a collection of data that specifies the number of replicas to maintain, the location of those replicas, the number of versions to maintain or the period of time to maintain versions of data, and the ability to control compression and encryption of data. This eliminates the over-replication problem

60409684.091002

associated with today's component model which protects data with RAID, volume management, network virtualization, backup, archiving, and block, file and application replication and snapshot solutions.

- h. A flexible data protection system that supports the modification of a protection policy to alter the options of encryption, replication, versioning, and compression. Additionally, the site's repositories for storing data can be changed as new sites open or existing sites close.
- i. A long-term data preservation system that can maintain data perpetually if the protection policy indicates that data must never be deleted. Data will be repaired automatically when data is found to be damaged during regular and automated integrity checking sweeps. This is in contrast to the many present-day problems associated with magnetic tape reliability over extended periods of time and the simple inability to know which tapes in a large repository have reliability problems without significant administrative effort.
- j. A cost-effective data protection system that replaces the traditional magnetic disk and magnetic tape systems with magnetic disks only. Magnetic tape has not kept pace with the more rapidly falling cost-per-gigabyte of magnetic disk technology. The following chart shows the trend line of raw media costs of magnetic disk drives vs. magnetic tape drives over the past decade.

60409684.091002



- k. A future-proof data protection system that eliminates the tape media obsolescence problem that occurs when tape drive technology advances. The system of the present invention does not suffer from these kinds of obsolescence problems since the disk node presents an Internet Protocol (IP) networking interface, not a disk-drive dependent interface like IDE, ATA, SCSI, or FibreChannel.
- l. A secure storage solution that maintains all enterprise data within the enterprise's internal infrastructure. By eliminating the need to store tapes at a third-party vaulting company, the potential of data loss due to theft or media mishandling is also greatly reduced.
- (3) In the area of data management, the present invention provides:
 - a. A greatly simplified data management system that eliminates tape media, tape drives, library units, offsite storage vaulting service providers, as well as backup, archiving and file replication servers and software. The administration of data protection is greatly reduced because multiple management products that would be required to control the existing collection of hardware and software are replaced by a single management interface.

60409684.091002

- b. A low-administration data management system that auto-discovers the presence of new storage capacity as each element is connected to the network.
- c. A relaxed, schedule service management model, enabled by the self-healing architecture. When hardware fails, it does not need to be replaced as soon as possible. A storage administrator is notified whenever a service action must be performed, but, since this invention has already automatically recovered from the failure, there is no need to replace components almost immediately after they fail. This relaxed service model is radically different from today's reactive, around-the-clock monitoring, near-immediate service/repair model.

The present invention further provides a system for the shared storage of computer-based data, comprising at least one client processor which forwards and receives computer data files; a plurality of storage nodes, at least one of the storage nodes being in communication with the at least one client processor; the storage nodes further being in communication with one another; and wherein at least one of the storage nodes receives computer data files from the at least one computer processor and distributes the contents in the data files to at least one other of the plurality of storage nodes.

The invention may further comprise replicating at least one of the computer data files received from the client processor to one other of the plurality of storage nodes so that there are at least two copies of the computer data files in the system. The invention may further comprise encrypting the computer data files. The invention may further comprise compressing the computer data files and storing at the storage nodes only those blocks of data that have changed over a predetermined time period.

Further the storage node may be comprised of at least one disk node for storing the computer data files and one to more port nodes to manage and distribute data received from the at least one client processor to the at least one disk node.

Further objects and advantages of this invention will become apparent from a consideration of the drawings and ensuing description.

60409684.091002

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates in schematic form the structure of a preferred embodiment of the present invention.

Figure 2 illustrates in schematic form a subset of the embodiment of Figure 1.

DETAILED DESCRIPTION

Figure 1 shows an illustrative embodiment of the invention, herein referred to as a storage grid. The embodiment of Figure 1 illustrates three sites (1), but it is to be understood that it can be expanded to enterprises with hundreds of sites. Within each site, zero or more disk nodes (2) and port nodes (3) are deployed.

Briefly, disk nodes (2) provide storage capacity to the storage grid. Each disk node may deliver substantial storage capacity (~1 terabyte initially) that is used for storing primary data as well as protecting the primary data that resides on other disk nodes.

Port nodes (3) provide the means for client computers to obtain access to the storage capacity across all disk nodes.

Port nodes and disk nodes are interconnected by a network (5) which is referred to herein as a storage grid network (5).

Figure 2 shows, for purposes of illustration only, the structure of a single disk node (2) and a single port node (3). A fully operational storage grid deployment may have many hundreds of each type of node. Software resides within each disk node and port node and manages the data storage, data protection and data management aspects of the storage grid.

The storage grid of Figure 1 logically may be seen to represent one consolidated enterprise-wide data storage and data protection system. Data may be stored across two or more sites to support the recovery of data in the event of a site disaster.

Within each site disk nodes (2), and port nodes (3) may be deployed. A particular site may have zero or more disk nodes and/or port nodes. The port node and disk node functions and operations are described below.

The storage grid network (5) allows disk nodes and port nodes to communicate with each other. Presently, many enterprise sites are already interconnected via IP-based metropolitan-area networks (MANs) or wide-area networks (WANs). The storage grid network (5) may be

60409684.091002

implemented as two or more site LANs that are interconnected via MAN or WAN routing devices.

Port nodes also connect to a customer's LAN (6) (local area network) to allow customer client systems (4) to store and retrieve storage grid data.

Port nodes (3) and disk nodes (2) may be implemented using server based products like IBM's NAS 100 server or Dell's 715N server product. The networking switches (8) and routers (7) may be implemented with common Ethernet connectivity components such as a Cisco 6500 switch or a Cisco 7400 router.

One or more client systems (4), shown in Figure 2, are in communication with each site. These clients may be well-known computer systems which are capable of functioning to request that data be written and read from the storage grid via the port node (3) interface.

Port nodes (3) may support both Sun's Network File System (NFS) and Microsoft's Common Internet File System (CIFS) file access interfaces (10).

NFS and CIFS operate over the port node's local filesystem (11). Local filesystems are provided by the operating system of the port node.

The operating system's filesystem supports a filesystem cache (12). The filesystem cache improves read performance by maintaining frequently accessed data requested from client systems on that port node. It also improves write performance by storing write requests locally to allow them to be updated more quickly as additional changes are made.

The port node's block-access device (24) holds the operating system for that system as well as the read cache.

A write-cache part of the filesystem cache (12) may be located on two or more block-access devices (21) of two or more disk nodes to prevent a loss of data in the event of a power failure to a single disk node or port node.

A Cache Manager (13) is responsible for obtaining files that are requested from the appropriate disk node system into a read-cache part of the file system cache (12). The Cache Manager is also responsible for moving least-recently referenced files out of the filesystem write-cache into one or more disk node systems.

60409684.091002

The Location Cache (16) may be queried by the cache manager (13) to determine whether the location of a requested file is already known to the port node queried. If the location cache (16) returns a negative response for a particular location request, the Location Manager (15) is queried to locate the specific disk nodes that contain a specific file. The Location Manager also determines the specific disk node(s) within the storage grid that new files are written to.

A port node Data Mover (18) communicates with one or more disk node Data Movers (19) to transfer files that have been written to the write-cache to their appropriate Disk node systems.

Disk node Data Movers (19) communicate with each other to re-replicate data whenever a hardware failure, such as a disk drive, occurs.

Port node and disk node data movers also provide a block-access interface (21) to allow the write cache and critical storage grid metadata to be stored on disk node systems.

The Protection Manager (14) maintains a protection policy for each filesystem (11). Each protection policy is defined by the storage grid storage administrator. The policy defines the number of replicas to be maintained, the desired location of these replicas, and the security, versioning and compression options for that filesystem.

For a new file that is being written to one or more disk nodes, the first version may be stored as a complete file. If a file has been updated, the port node Version Manager (17) asks the Protection Manager (14) about the versioning policy for that file. The Protection Manager may have policies set on the maximum number of versions to maintain, the maximum period to retain versions, and whether or not to perform delta-compression on older versions of a file. Based on this information, the port node Version Manager (17) instructs the various disk node Version Managers (20) that contain previously stored versions of the updated file to perform file removal operations if the file count or the file retention period that was specified in the protection policy is exceeded, or it may delta compress older versions if that protection policy option is enabled.

Each disk node and each port node system includes a Node Manager (23). The Node Manager may provide information on the assets, configuration, status, capacity, and performance of each node in the storage grid. This information allows each administrator to view its

60409684.091002

resources as well as the other resources within their enterprise that are part of their storage grid. This information may also be used by the location manager to identify space availability when deciding on where to place new or updated files.

Operation of the Embodiments of Figures 1 and 2

An illustrative example of the operation of the embodiments of the present invention will now be described. Figure 1 displays an exemplary storage grid system and is shown as having three sites (1). Within each site, there are one or more disk nodes (2) that provide storage capacity to the storage grid. Within each site, there are also zero or more port nodes (3) to allow client systems to store and retrieve data from the storage grid's disk nodes.

Every client-created file on a storage grid may be grouped with other files into a filesystem (11). Each filesystem in Figure 2 is associated with a protection policy set by a Protection Manager (14) in Figure 2 that defines, for all of the files within that filesystem, the number of replicas of data desired and the desired site location for those replicas.

When a file is written from a client system (4) to one of the port nodes, the port node software is responsible for replicating data to any of the disk nodes within the site(s) that were specified in the protection policy. This is illustrated in the schematic drawing in Figure 2 where clients (4) are shown connected to filesystem (11). The port node maintains a mapping of files in its location cache (16) and their specific disk node location information. When a file is requested to be read, it can be acquired from any Disk node that has a replica copy of the same data.

When a file is requested from a client via the port node, the storage grid software, by default, may be programmed to access that file from the disk node that is closest to the requesting port node in terms of geographical distance. If any of the disk nodes that are being referenced are not operational at the time of a port node request, the data is delivered from any of the surviving disk node systems at other sites. This is made to occur automatically by programming a hierarchy of search techniques for use with each grid node and is transparent to the client's application.

60409684.091002

The following scenarios will be described in this section:

(a) Creation of a new file; (b) Updating an existing file; (c) Reading a file; (d) Deleting a file; (e) Re-replicating data from a failed drive; (f) Retention purging; (g) Port node failover; (h) Drive Failure Service Model; and (i) Self-healing storage;

(a) Creating a new file: One of the ways a new file can be created and stored in the storage grid will now be described. A client system (4) application can create a new file by writing it to a port node NFS/CIFS daemon (10) via a series of successive NFS or CIFS write operations. These NFS/CIFS write operations are forwarded to the local filesystem (11). The new file remains in the write filesystem cache (12) until it is flushed. At that time, the Protection Manager (14) provides the protection policy data to allow the Location Manager (15) to determine the number of replicas of that file the user or administrator has chosen to maintain and the specific disk nodes where those file replica(s) should be stored. The port node Data Mover (18) is responsible for initiating the write of the file replicas from the write cache to the disk node nodes that were identified by the Location Manager, wherever those nodes may be located on the storage grid network. When the writes are complete, the Location Cache (16) is updated on the local port node to expedite subsequent read, update or delete requests for that file.

(b) Updating an existing file: When a file already exists in the storage grid and is being updated by the client system application (4), it is maintained in the write filesystem cache (12) until it is flushed. At that time, the port node Version Manager (17) queries the Protection Manager about the versioning policy for that file. The Protection Manager may have policies set on the maximum number of versions to maintain, the maximum period to retain versions, and whether or not to perform delta-compression on older versions of a file. Based on this information, the port node Version Manager (17) instructs the various disk node Version Managers (20) that contain previously stored versions of the updated file to perform file removal operations if the file count is exceeded, or the file retention period is exceeded or it may delta compress older versions if that policy option is enabled.

(c) Reading a file: One way that a file could be read will now be described. When the client system application (4) requests a read of a

60409684.091002

file from its port node (3), the file may already be in its local read cache. If this is the case, the read operation is satisfied quickly. If the file is not contained in the local read cache, the Location Cache (16) is queried by the cache manager (13) to determine whether this Port node already knows the location of a specific file. If the location cache currently contains the information to identify the collection of disk nodes where the requested file resides, a Data Mover (18) operation is set up between the disk node that contains the data and the read cache of the port node that requested the data. If the location cache does not currently contain the information to identify the collection of disk nodes where the requested file resides, the Location Manager (15) is responsible for identifying all of the disk nodes that have a replica of the requested file. Once it has been verified that the disk nodes actually have the requested file, the Location Cache (16) is updated with that information and a data mover operation between one of the disk nodes that has the file and the requesting port node system's filesystem read cache takes place.

(d) Deleting a file: One of the ways deletion of files can be accomplished will now be described. When a request for file deletion occurs, the protection policy for a filesystem determines the action which may be taken. The options are to: immediately delete the file without administrative confirmation; to delete a file only after an administrator has confirmed the delete; and to honor the protection policy's retention period over client application delete requests and not delete the file if it is still within the period of retention. Other options include conditionally deleting replicas from specific sites, and deleting all previous versions as well as the latest version. When a file is actually deleted, the Location Manager is queried to determine all of the disk node locations where that file may be located. The delete operation is carried out by one or more disk node Version Managers (20). The entries previously stored in the location cache are eliminated as well.

(e) Re-replicating data from a failed drive: When a disk drive fails in a disk node, all of the files it contained have already been replicated elsewhere on the storage grid network. The task, then, is to locate and engage those disk nodes that contain a replica of the lost files to re-replicate their content to other disk node systems at the

60409684.091002

site of the disk node system that experienced the drive failure. This is performed by the disk nodes' Node Manager (23).

(f) Retention Purging: The protection policy for a filesystem may include a time-period based retention component. An example of a time-based retention policy would be to keep files for 7 years after their data of creation to comply with the Securities and Exchange Commission's regulatory rules on corporate data retention. The finest granularity for time-period retention is a day. Once a day, a retention purge is scheduled in order to remove files whose retention period has elapsed. This is accomplished by disk nodes' Node Manager (23). Before these files are actually deleted, they may optionally be presented to the storage administrator to confirm the delete operation.

(g) Port node failover: When a port node failure occurs, a new port node is physically attached to the client system network and to the storage grid network. This is attached by a systems administrator. Once the port node's software is installed, the new port node is configured by the system administrator to assume the identity of the failed port node. The new port node is provided with all of the information necessary to take over for the failed port node since all of the necessary port node metadata is maintained within the storage grid.

(h) Drive Failure Service Model: Another feature of the storage grid of the present invention is a relaxed scheduled service model. For example, a disk node may have four disk drives. As each of these drives fails, the data that was lost is automatically re-replicated to other disk nodes by the surviving disk nodes. The failed drive remains in the disk node. When all four drives have failed, the entire disk node is replaced. This is an improvement over the present technology, in which the failed drive is replaced more or less immediately upon its failure.

(i) Self-healing storage: When a new file is or an updated file is written into a port node, and then into a disk node, an algorithmic checksum is computed that can be used to detect changes to that file over time. Periodically, each disk nodes' Node Manager is responsible for verifying the integrity of each file that it contains by recalculating a checksum and comparing it with the checksum that was computed when the file was first written to the storage grid. If the two checksums are different, the Node Manager (23) that is running in that disk node is responsible locating a good copy of the file

60409684.001002

elsewhere in the storage grid and replacing the bad file with the known good copy of the file.